Gas dynamic analogous exposure approach to interaction intensity in multiple-vehicle crash analysis: Case study of crashes involving taxis

Fanyu Meng[a], Wai Wong[a], S.C. Wong[a,1,*], Xin Pei[b,2], Y.C. Li[a], Helai Huang[c]

[a.] *Department of Civil Engineering, The University of Hong Kong, Hong Kong, China*

[b.] *Department of Automation, Tsinghua University, Beijing, China*

[c.] *School of Traffic and Transportation Engineering, Central South University, Changsha, China*

Abstract

Exposure is a frequency measure of being in situations in which crashes could occur. In modeling multiple-vehicle crash frequency, traditional exposure measures, such as vehicle kilometrage and travel time, may not be sufficiently representative because they may include situations in which vehicles rarely meet each other and multiple-vehicle crashes can never happen. The meeting frequency of vehicles should be a better exposure measure in such cases. This study aims to propose a novel Gas Dynamic Analogous Exposure (GDAE) to model multiple-vehicle crash frequency. We analogize the meeting frequency of vehicles with the meeting frequency of gas molecules because both systems consider the numbers of the meetings of discrete entities. A meeting frequency function of vehicles is derived based on the central idea of the classical collision theory in physical chemistry with consideration of constrained vehicular movement by the road alignments. The GDAE is then formulated on

*Corresponding author.
E-mail addresses: fymeng91@gmail.com (F. Meng), waiwong1012@connect.hku.hk (W. Wong), hhecwsc@hku.hk (S.C. Wong), peixin@tsinghua.edu.cn (X. Pei), joeyliyc@connect.hku.hk (Y.C. Li), huanghelai@csu.edu.cn (H. Huang).
[1] Fax: +852 2559 5337.
[2] Fax: +86 1062795043.

the basis of the major factors that contribute to the meeting frequency of vehicles. The proposed GDAE is a more representative proxy exposure measure in modeling of multiple-vehicle crash frequency because it further investigates and provides insight into the physics of the vehicle meeting mechanism. To demonstrate the applicability of the GDAE, zonal crash frequency models are constructed on the basis of multiple-vehicle crashes involving taxis in 398 zones of Hong Kong in 2011. The GDAE outperforms the conventional time exposure in multiple-vehicle crash modeling. To account for any unobservable heterogeneity and to cope with the over-dispersed count data, a random-parameter negative binomial model is established. Explanatory factors that contribute to the zonal multiple-vehicle crash risk involving taxis are identified. The proposed GDAE is a promising exposure measure for modeling multiple-vehicle crash frequency.

Keywords: Gas dynamic analogy, Exposure, Multiple-vehicle crash frequency, Zonal crash frequency, Taxi safety

1 Introduction

In road safety, crash frequency modeling is an important and useful tool for identification of factors that contribute to crash frequency. Remedy measures or policies can be formulated and implemented on the basis of the identified factors to enhance road safety. Depending on the purpose of the given study, crash frequency models in terms of different categories, such as the sites of interest (e.g., intersections, road segments, highways, and zonal networks), the numbers of vehicles involved (e.g., single and multiple vehicles), the vehicle types (e.g., motorcycles, taxis, and trucks), and injury severity (e.g., slight-injury and killed or seriously injured), can be established.

1.1 Exposure to crash risk

Exposure measures are the essential elements that are tightly linked to all kinds of crash frequency models. Exposure has been defined differently over the decades. Chapman (1973) defined exposure as the number of opportunities for crashes of a certain type to occur over a given time in a given area. Wolfe (1982) later offered a modified definition of exposure as simply being in a situation that incurs some risk of being involved in a crash and expressed risk as the number of crashes that take place in the same situation in a certain period divided by exposure. More recently, Elvik (2015) defined an event-based definition of exposure in which each event with the potential to generate a crash is interpreted as a trial, as defined in probability theory. Although certain levels of differences lie in these definitions, they all serve the single purpose of determining crash risks or accident rates that indicate the relative risk levels of various traffic situations (Wolfe, 1982).

Broadly speaking, the exposure measure is rather conceptual, and direct measurement may not be feasible in many situations. In practice, although the use of exposure measures is constrained by the availability and quality of data (Naci et al., 2009), various proxy measures have been developed and used in different crash frequency analyses, including population and fuel consumption (Amoh-Gyimah et al., 2017; Fridstrøm et al., 1995), traffic volume (Chiou and Fu, 2015; Heydari et al., 2017; Qin et al., 2004, 2006; Wong et al., 2007), travel time (Chipman et al., 1993; Imprialou et al., 2016), vehicle-miles traveled (Li et al., 2003; Pei et al., 2016), potential conflict counts (Bie et al., 2005; Wong et al., 2006), and quasi-induced exposure (Huang and Chin, 2009; Jiang et al., 2014; Stamatiadis and Deacon, 1997).

In general, zonal-level exposure measures such as population are suitable for zonal crash frequency models, and micro-level exposure measures such as traffic flows are more

72 frequently used in modeling crash frequencies at specific roadway entities such as road
73 segments and junctions. For instance, Lee et al. (2015) used zonal population as exposure
74 measure to develop macroscopic multivariate crash analysis reporting models. It was
75 anticipated to efficiently help policymakers allocate resources to improve road safety for
76 different zones. Similarly, Amoh-Gyimah et al. (2017) incorporated population and vehicle-
77 kilometers in a macroscopic crash model and investigated the effects of spatial variations in
78 the unobserved heterogeneity. The results showed that when the spatial variability is
79 considered, an increase in the population of young people increased the crash risk, although
80 the parameter of this variable was negative. For crash risk at road segments, Pei et al. (2012)
81 estimated the travel distance and travel time across 112 road segments in Hong Kong using
82 global positioning system (GPS) data and investigated the influence of these two exposure
83 measures on the relationship between speed and crash risk. Their results revealed a positive
84 correlation between the average speed and crash risk when the distance exposure was adopted.
85 In contrast, average speed had a negative correlation to the crash risk when the time exposure
86 was used. Tulu et al. (2015) investigated pedestrian crash frequency for two-way two-lane
87 rural roads in Ethiopia by considering the product of vehicle volume and pedestrian volume
88 as the exposure measure and established a random-parameter negative binomial model. A
89 nonlinear effect of the exposure measure was found, and the modeling results indicated that
90 the proportion of the daily crossing volume by pedestrians younger than 19 years of age
91 could be used to explain pedestrian exposure in further studies. However, these exposure
92 measures are highly aggregated measures that may not adequately represent exposure to crash
93 risk. For instance, a greater zonal population is not necessarily equivalent to a greater number
94 of commuters, and a greater number of commuters does not mean that all of them are
95 exposed to situations that could possibly develop into a crash (e.g., a pedestrian walking on a

96 street without any vehicles). Similarly, Qin et al. (2006) also pointed out that the
97 conventional aggregated exposure measures do not account for temporal variations in traffic.
98
99 Because different types of crashes have different causes, exposure to these traffic hazards
100 (crash risk) may vary. To better identify the factors that contribute to the crash risk, it is of
101 great importance to use a more representative exposure measure for the model development.
102 Many researchers attempted to formulate different kinds of exposure measures by using
103 disaggregated data and considering the mechanism for a potential crash. In a study
104 concerning crash rate prediction in two-lane highway segments, Qin et al. (2004) formulated
105 different exposure functions for single-vehicle crashes and multiple-vehicle crashes in three
106 directions: the same direction, opposite directions, and intersecting directions. The
107 disaggregated flow for each direction of the highway and the segment length were used for
108 the formulations. The results showed that most of the proposed exposure functions had linear
109 relationships with the crash frequency of their corresponding crash types, whereas the
110 conventional exposure measure, vehicle-miles traveled, had nonlinear relationships with the
111 crash frequencies. This finding revealed that their proposed exposure functions would be
112 more representative than vehicle-miles traveled in these scenarios. Instead of using hourly
113 traffic volume, Miranda-Moreno et al. (2011) applied disaggregated flows by movement type
114 and vehicle type in their study of crash risk at intersections. They proposed that the
115 movement types exhibited by vehicles and bicyclists at an intersection may have different
116 effects on the crash risk. Disaggregated flows were used to formulate three exposure
117 measures: aggregated flows, motor vehicle flows aggregated by movement type, and
118 potential conflicts between motor vehicles and cyclists. The products of the different
119 combinations of conflicting disaggregated flows were considered to indicate the conflicting
120 volumes. Similar concepts have been included in a more advanced model—the latent class

model with Bayesian inference—to study the unobserved heterogeneity in pedestrian and cyclist crashes (Heydari et al., 2017).

Multiple-vehicle crashes are one of the important crash types in which transport authorities have great interest. For instance, a concerned local authority may wish to identify the factors that contribute to the risk of multiple-vehicle crashes involving trucks and private cars for policy formulations. The amount of energy released in a crash involving a truck could be huge, and the private car driver and passengers could be seriously injured or killed due to the great size difference between the two vehicles. Chen and Xie (2016) studied the role of average annual daily traffic (AADT) in the prediction of multiple-vehicle crash frequency by establishing generalized additive models and piecewise linear negative binomial regression models. Forty-eight three-approach signalized intersections and 52 four-approach signalized intersections were included and modeled separately; the results revealed that a nonlinear functional form of AADT performed better than a linear form in multiple-vehicle crash frequency models. However, conventional exposure measures that are normally adopted for multiple-vehicle crashes may not be sufficiently representative, because they may include situations in which vehicles rarely meet and multiple-vehicle crashes can never happen. Because multiple-vehicle crashes can only happen when vehicles meet, their meeting frequency should be a more representative exposure measure in these cases.

1.2 Methodological challenges in crash modeling

With advancements in modeling methods, recent crash frequency models have been established to address various important issues, such as cross-equation error correlation, crash frequency by injury severity, unobserved heterogeneity, and space- and time-specific heterogeneity, which has enabled more accurate estimation of the relationships between crash

146 frequency and various contributive factors. The cross-equation error correlation naturally
147 arises from unobserved factors that may affect multiple crash counts or the injury levels of
148 different types of crashes (Serhiyenko et al., 2016), different occupants in the same crash
149 (Russo et al., 2014), or different crash severity levels (Anastasopoulos, 2016; Sarwar and
150 Anastasopoulos, 2017) simultaneously, or from the temporal correlation at the same road
151 entity (Mannering et al., 2016). Multivariate modeling approaches have been shown to
152 adequately address cross-equation error correlation and to outperform their univariate
153 counterparts in multiple studies (Barua et al., 2015; Huang et al., 2017; Serhiyenko et al.,
154 2016). In addition to cross-equation correlation, unobserved heterogeneities across various
155 road entities, various periods, or both are also worthy of note; if not addressed, they may
156 cause problematic estimation results by introducing variation in the effects of observed
157 variables (Mannering et al., 2016). The most common approach to consider full unobserved
158 heterogeneities in crash likelihood modeling is a random-parameter model, which has been
159 thoroughly investigated in various studies (Anastasopoulos and Mannering, 2009; Barua et al.,
160 2016; Bhat et al., 2014; Chen and Tarko, 2014; Coruh et al., 2015; Venkataraman et al., 2011;
161 Venkataraman et al., 2013). In addition, the latent-class model is another possible way to
162 model unobserved effects in crash data (Buddhavarapu et al., 2016; Heydari et al., 2016), and
163 random parameters can be further adopted within each class (Xiong and Mannering, 2013).
164 Moreover, the consideration of space- and time-specific heterogeneity and spatial/temporal
165 correlation has provided new insights for scholars investigating crash frequency modeling
166 (Chiou et al., 2014, 2015; Huang et al., 2017).
167
168 Furthermore, some studies have incorporated heterogeneous and/or space-time effects in
169 exposure measures, where AADT is a preferable exposure measure in modeling crash risk
170 when considering spatial heterogeneity or spatial correlation. Barua et al. (2016) established a

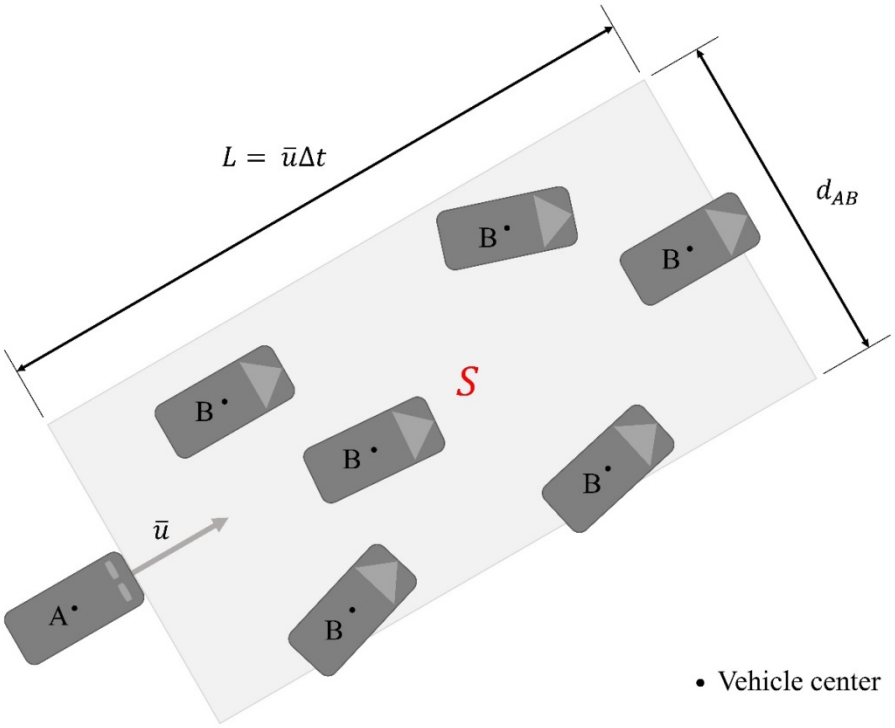
multivariate random-parameter model for severe and no-injury collisions in Vancouver and showed that the exposure variable contained spatial heterogeneity. Similar results were found by Huang et al. (2017) in a multivariate Poisson log-normal model with spatial random effects. Moreover, Chiou and Fu (2015) modeled the spatiotemporal dependence of the crash frequency and severity and concluded that temporal effects were more suitable for crash frequency than for crash severity because the temporal effects mainly came from the traffic volume, which was closely correlated with the crash frequency. Kroyer et al. (2016) studied the effect of pedestrian and bicyclist flows on intersection crash frequencies, in which the temporal variability of the exposure effects was considered with the use of an exposure distribution curve. Safety performance functions were proposed in relation to the increased model reliability achieved with short observational periods. Although some studies have considered the temporal effects of the exposure measures on the risk of multiple-vehicle crashes, few studies have considered the development of the exposure measures from the perspective of their meeting mechanisms, which could yield a more representative measure.

In this paper, we propose a gas dynamic analogous exposure (GDAE) to model multiple-vehicle crashes. The meeting frequency of vehicles is analogized with the meeting frequency of gas molecules, as both systems describe the number of meetings of discrete entities. The meeting frequency function of vehicles that further considers the mechanism of their meeting frequency is derived based on the central idea of the classical collision theory in physical chemistry. The GDAE is formulated on the basis of the major identified factors with correction terms. Negative binomial (NB) models with only an exposure variable are established for multiple-vehicle crashes, in which the GDAE is compared with the traditional travel time exposure measures, using the data of crashes involving taxis in Hong Kong in 2011 as a case study. The results reveal that the GDAE performed better than the traditional

196 travel time exposure measures for multiple-vehicle crashes. Thus, multiple-vehicle crash
197 frequency models are established using the GDAE and other potential explanatory variables
198 that contribute to the crash risk. A random-parameter negative binomial (RPNB) model is
199 used to account for the unobserved heterogeneity in the dataset. Influential factors with a
200 significant association with the risk of multiple-vehicle crashes involving taxis in Hong Kong
201 are identified.
202
203 The remainder of this paper is organized as follows. In Section 2, the meeting frequency
204 function and the GDAE are derived, and the methods of modeling crash data with
205 consideration of the presence of heterogeneity are discussed. Section 3 presents the
206 background, databases, results, and discussions regarding the case study of modeling
207 multiple-vehicle crashes involving taxis in Hong Kong. Section 4 provides concluding
208 remarks and recommendations for future research.
209
210 ## 2 Methods
211 This section first derives the meeting frequency function and the GDAE. The GDAE is a
212 potentially more representative proxy measure of exposure for modeling multiple-vehicle
213 crash frequency because it provides further insight into the physics of the vehicle meeting
214 mechanism. The modeling methods of crash data in the form of panel data with consideration
215 of the existence of overdispersion and heterogeneity are then presented.
216
217 ### 2.1 Meeting frequency function and GDAE
218 The meeting frequency of vehicles can generally be analogized with the meeting frequency of
219 gas molecules because both systems consider the meeting quantities of discrete entities. The
220 major difference between the two systems is that molecules move freely and randomly,

whereas vehicular movements are constrained by road alignments. In this subsection, the meeting frequency function of vehicles is first derived by leveraging the concept of the classical collision theory in physical chemistry (Laidler, 1973). However, it should be stressed that a meeting of molecules, which is usually called a collision of molecules in physical chemistry, only corresponds to a meeting of vehicles, but not a crash. The resultant meeting frequency function offers physical insight into the meeting mechanisms of multiple vehicles. The GDAE is then formulated by the identified factors that contribute to the meeting quantities.

Because vehicles interact with surrounding vehicles as they travel, their speeds should be similar and can generally be assumed to follow a distribution with mean $\bar{u}$ (i.e., $\overline{|\boldsymbol{u}|} = \bar{u}$). Consider a vehicle A traveling with a mean speed, $\bar{u}$, as shown in Fig. 1. d_{AB} is a conceptual effective meeting width that depends on various factors, such as the sizes of the type A and B vehicles and the characteristics of the road segment. For instance, if vehicle A is traveling in the middle lane of a three-lane road, d_{AB} is approximately equal to the width of the three-lane road. However, if vehicle A is traveling on a two-lane road, d_{AB} is at most equal to the width of a two-lane road. In a given time interval, Δt, the distance traveled by vehicle A is $L = \bar{u}\Delta t$, and the influential area swept over by $d_{AB}$is given by $S = Ld_{AB}$.



239

240 Fig. 1. Idealized scenario of meetings of type A vehicle and type B vehicles within influential
241 area S in given time interval Δt.

242

243 Denote the average number density of type B vehicles by $n_B = N_B/R$, where N_B is the
244 average number of type B vehicles in a given time interval, Δt, and a given road space, R.
245 Imagine that the type B vehicles with their centers lying in S are stationary, as shown in
246 Figure 1. The number of type B vehicles met by vehicle A in the time interval, Δt, is given by
247 $d_{AB}\bar{u}n_B\Delta t$. However, type B vehicles are not really stationary, thus the mean speed, $\bar{u}$,
248 should be replaced by the mean relative speed, $\overline{|\boldsymbol{u_r}|}$ or $\bar{u}_r$. Consider the relative velocity
249 vector of any pair of type A and B vehicles as illustrated in Fig. 2. $\boldsymbol{u_A}$ and $\boldsymbol{u_B}$ are the velocity
250 vectors for vehicles A and B. θ is the angle between $\boldsymbol{u_A}$ and $\boldsymbol{u_B}$.

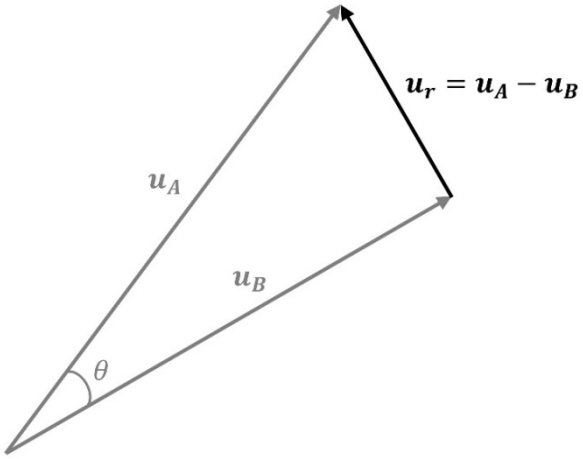


Fig. 2. Relative velocity of any pair of type A and B vehicles.

Because $|\boldsymbol{u_r}|^2 = \boldsymbol{u_r} \cdot \boldsymbol{u_r}$ and $\boldsymbol{u_r} = \boldsymbol{u_A} - \boldsymbol{u_B}$,

$$|\boldsymbol{u_r}|^2 = |\boldsymbol{u_A}|^2 - 2cos\theta|\boldsymbol{u_A}||\boldsymbol{u_B}| + |\boldsymbol{u_B}|^2$$

Taking average on both sides,

$$\overline{|\boldsymbol{u_r}|^2} = \overline{|\boldsymbol{u_A}|^2} - 2\overline{cos\theta|\boldsymbol{u_A}||\boldsymbol{u_B}|} + \overline{|\boldsymbol{u_B}|^2} \quad (1)$$

In the classical collision theory, the second term on the right-hand side of Eq. (1) is zero (i.e., $\overline{\boldsymbol{u_A} \cdot \boldsymbol{u_B}} = 0$), because gas molecules move freely and randomly and the average case appears to be $\pi/2$ meeting angle. However, this is not the case in the meeting frequency of vehicles because vehicular movements are constrained by road alignments. Using first-order Taylor series approximation at ($\overline{|\boldsymbol{u_A}|}$, $\overline{|\boldsymbol{u_B}|}$, $\overline{cos\theta}$),

$$\overline{|\boldsymbol{u_r}|}^2 \approx \overline{|\boldsymbol{u_A}|}^2 - 2\overline{cos\theta}\ \overline{|\boldsymbol{u_A}|}\ \overline{|\boldsymbol{u_B}|} + \overline{|\boldsymbol{u_B}|}^2$$

Because $\overline{|\boldsymbol{u_A}|} = \overline{|\boldsymbol{u_B}|} = \bar{u}$,

$$\bar{u}_r \approx \sqrt{2 - 2\overline{cos\theta}}\ \bar{u}$$

The angle between $\boldsymbol{u_A}$ and $\boldsymbol{u_B}$, θ, can be assumed to follow a probability distribution denoted by $f(\theta), \forall\theta \in (-\pi, \pi]$, which should be system-dependent. For instance, if a network has many junctions or the local drivers frequently overtake each other, the probabilities of θ at larger values could be higher. Nevertheless, because networks are designed to segregate

traffic traveling in different directions, cases such as head-to-head vehicle meetings (i.e., $\theta \approx \pi$) are usually infrequent. Therefore,

$$\overline{cos\theta} = \int_0^{2\pi} cos\theta \, f(\theta) \, d\theta$$

$\exists$ an unknown constant effective meeting angle $\theta^* \in [0, 2\pi)$ $s.t. cos\theta^* = \overline{cos\theta}$. Thus, $\bar{u}_r$ can be rewritten as

$$\bar{u}_r \approx \sqrt{2 - 2cos\theta^*}\, \bar{u}.$$

Given that the average number density of type A vehicles is $n_A = N_A/R$, where N_A is the number of type A vehicles in a given time interval, Δt, and a given road space, R, the meeting frequency of type A and B vehicles, m_{AB}, in the given time interval, Δt, and the given road space, R, is given by Eq. (2):

$$m_{AB} = d_{AB}\, \bar{u}_r\, n_A\, n_B\, R\, \Delta t. \tag{2}$$

Because $\bar{u}_r \approx \sqrt{2 - 2cos\theta^*}\, \bar{u}$, $n_A = N_A/R$ and $n_B = N_B/R$,

$$m_{AB} \approx \sqrt{2 - 2cos\theta^*}\, d_{AB}\, \bar{u}\, \frac{N_A}{R} \frac{N_B}{R}\, R\, \Delta t. \tag{3}$$

In addition, using the definitions of N_A and N_B, the total travel time of type A and B vehicles, T_A and T_B, can be expressed as Eq. (4a) and (4b), respectively.

$$T_A = N_A \Delta t \tag{4a}$$

$$T_B = N_B \Delta t \tag{4b}$$

Substituting Eq. (4a) and (4b) into Eq. (3),

$$m_{AB} \approx \frac{\sqrt{2-2cos\theta^*}\, d_{AB}}{\Delta t} \frac{\bar{u}}{R}\, T_A T_B = C I T_A T_B, \tag{5}$$

where $C = \sqrt{2 - 2cos\theta^*}\, d_{AB}/\Delta t$ is an unknown constant for a given Δt; and $I = \bar{u}/R$ is a state-topological factor that captures both the operation state (i.e., $\bar{u}$) and the road space (i.e., R) of a network. In particular, if the meeting frequency for the same vehicle type is

280 considered (i.e., type A = type B), the corresponding meeting frequency function is given by
281 Eq. (6),

$$m_{AA} \approx C' I {T_A}^2, \tag{6}$$

282 where $C' = C/2$ is also an unknown constant. The factor of $1/2$ is introduced to avoid
283 double-counting the meeting of the same pairs of vehicles. Therefore, more generically, the
284 GDAE is applicable to multiple-vehicle crashes but not simply multiple types of vehicles.
285 The derived meeting frequency function provides a theoretical foundation for quantifying
286 exposure in multiple-vehicle crashes and offers insights into the physics of the vehicle
287 meeting mechanism by revealing the physical quantities that govern the number of meetings.
288 The meeting frequency function links the meeting quantity with the effective meeting angle
289 and the width, mean speed, road space, and total travel time of type A and B vehicles. With
290 all other factors kept constant, the meeting quantity should increase with the mean speed of
291 the vehicles for a given road space and time period because the area of influence covered by
292 the vehicles increases with their mean speed in the spatiotemporal volume, leading to a
293 greater likelihood of meeting. Similarly, the meeting frequency should increase with the total
294 travel time of the type A and B vehicles.
295
296 Compared with conventional exposure measures, it should be a more representative proxy
297 measure for exposure in multiple-vehicle crashes because it further explores the mechanism
298 of such meetings. However, direct evaluation of Eq. (5) may not be possible because C
299 comprises two unknown constants, d_{AB} and θ^*. Nevertheless, the function identifies the
300 major factors, I and $T_A T_B$, related to the meeting frequency of two types of vehicles. In
301 practice, instead of directly applying the meeting frequency function, an alternative proxy
302 exposure measure, the GDAE, as shown in Eq. (7) should be adopted because it is formulated

303 by the identified factors with data inputs that can be readily extracted from different
304 databases.

$$GDAE = I^{\gamma_1}(T_A T_B)^{\gamma_2}, \quad (7)$$

305 where γ_1 and γ_2 are the correction terms of the GDAE that account for any distinctions
306 between the idealized scenario and the reality. Because γ_1 and γ_2 are the model parameters to
307 be calibrated, both the GDAE and the sensitivity parameters of the explanatory variables are
308 calibrated simultaneously upon regressions of the crash frequency models. Moreover, the
309 calibrated correction terms should be positive because both *I* and $T_A T_B$ should increase with
310 m_{AB}, as shown in Eq. (5).
311

312 **2.2 Crash frequency modeling**

313 Because overdispersion exists in most crash data, NB regression is more favored in crash
314 frequency modeling than the use of the Poisson model because it is commonly used to deal
315 with overdispersion (Coruh et al., 2015). Moreover, unobserved heterogeneity may lead to
316 underestimated standard errors associated with the estimated coefficients and thus inflated *t*-
317 ratios (Venkataraman et al., 2013). Therefore, an RPNB model was used in this study to
318 better address the overdispersion and the unobserved heterogeneity in the crash dataset. The
319 probability that y_i crashes occur in zone i is as follows:

$$P(y_i) = \frac{\Gamma\left(\frac{1}{\alpha} + y_i\right)}{\Gamma\left(\frac{1}{\alpha}\right) y_i!} \left(\frac{\frac{1}{\alpha}}{\left(\frac{1}{\alpha}\right) + \lambda_i}\right)^{\frac{1}{\alpha}} \left(\frac{\lambda_i}{\left(\frac{1}{\alpha}\right) + \lambda_i}\right)^{y_i}, \quad (8)$$

320 where λ_i represents the expected number of crashes in zone i in a certain period, and α is the
321 overdispersion parameter. The Poisson regression model is a special case of the NB
322 regression model, in which α approaches zero.
323

If the crashes are divided into I zones and T periods, the expected number of crashes in zone i and period t can be determined by a series of explanatory variables in log-link form, as follows:

$$\lambda_{it} = E_{it}\,\mathrm{EXP}(\boldsymbol{\beta}\mathbf{X}_{it} + \epsilon_i), \tag{9}$$

where E_{it} is the exposure measure, $\boldsymbol{X}_{it}$ represents the vector of the explanatory variables, $\boldsymbol{\beta}$ is the vector of the coefficients to be estimated, and $\mathrm{EXP}(\epsilon_i)$ is the error term, which follows a gamma distribution with a mean of 1 and variance α (Washington et al., 2011). The random parameters of RPNB models are generally stated in the form of a mean and a random term as follows (Greene, 2007; Mannering et al., 2016):

$$\beta_i = \beta + \varphi_i, \tag{10}$$

where β_i is the estimable coefficient for the ith zone, β is the fixed proportion of the random parameter, and φ_i is a normally distributed parameter among various zones with mean 0 and variance ${\sigma_i}^2$. The parameter β_i is considered to be random only if the variance ${\sigma_i}^2$ of the random part φ_i is greater than zero (Anastasopoulos and Mannering, 2009). To identify the explanatory factors that contribute to the crash risk and determine whether their effects are heterogeneous, Z-tests should be performed to determine the significance levels of the estimated coefficients.

In particular, if the multiple-vehicle crash frequency is modeled, the GDAE proposed in the previous subsection should be considered as one of the candidate exposure measures, because it further investigates the mechanism of vehicle meeting and is anticipated to be a more representative proxy measure for exposure in these cases.

3 Case Study: Taxi Crashes

The proposed GDAE is expected to be a more representative proxy measure of exposure for modeling multiple-vehicle crash frequency. However, such an anticipation lacks statistical evidence and empirical data support. To demonstrate the applicability of the proposed GDAE, zonal multiple-vehicle crash frequency models were developed on the basis of crashes involving taxis in 398 zones of Hong Kong in 2011. Taxis were chosen as the vehicle type of interest for the case study because they are generally regarded as a risky group in road safety (i.e., type A vehicles, taxi; type B vehicles, all other types of vehicles). The explanatory factors that contribute to the multiple-vehicle crash risk were identified.

3.1 Taxi safety

Compared with nonprofessional drivers, professional road users such as taxi drivers and bus drivers are considered to face a greater risk of involvement in crashes, especially fatal ones, because their exposure to this risk is higher (Baker et al., 1976, Johnson et al., 1999). According to the Transport Department of Hong Kong, the involvement of taxis in road accidents has increased over the past decade, ranking second among all vehicular classes in 2013, in which 4300 taxis were involved in crashes (Transport Department, 2014). Table 1 reveals the crash risk comparison between taxis and all vehicle types for crashes with different levels of injury severity (i.e., slight injury versus killed or severely injured) based on crash data in 2011 in Hong Kong. (A detailed database description is covered in Section 3.2.) The crash risk here was defined as the number of crashes per million vehicle hours. The results showed that taxis generally were more likely to be involved in crashes, especially slight-injury crashes, than vehicles overall. Therefore, taxis were the chosen vehicle type for this case study.

370 Table 1. Crash risk comparison.

	Taxi crash risk (crash/million vehicle hours)	Total traffic crash risk (crash/million vehicle hours)
Slight-injury crashes	26.81	18.87
KSI crashes	3.53	3.31
Crashes of all types	30.34	22.18

371 KSI, killed or severely injured.

372

373 In the past decade, preliminary studies of taxi safety have been conducted with several points
374 of focus. First, studies have examined taxi drivers' views of the factors responsible for their
375 risky behavior and have used questionnaire surveys as the optimal means of eliciting these
376 views (Machin and De Souza, 2004; Rosenbloom and Shahar, 2007; Shams et al., 2011). In
377 addition to collecting and analyzing data on taxi drivers' attitudes, some researchers have
378 objectively analyzed their behavior using observed data and identified a tendency toward
379 aggression (Burns and Wilde, 1995; Dalziel and Job, 1997; Sullman et al., 2013). As data
380 emerged on the aggression of taxi drivers relative to other road users, taxi-safety researchers
381 began to consider the quantifiable relationship between crash risk and some influential factors
382 (Lam, 2004; La et al., 2013).

383

384 Although a few studies have focused on identifying the explanatory factors that contribute to
385 the crash risk of taxis, the adoption of an exposure measure is still restricted to traditional
386 travel time and travel distance exposure. However, these measures may not be sufficiently
387 representative for multiple-vehicle crashes involving taxis. A more appropriate and
388 representative exposure measure could facilitate the identification of contributory factors that
389 influence crash risk.

390

3.2 Data

This subsection presents the preparation of a comprehensive dataset that comprises necessary information for modeling zonal multiple-vehicle crash frequency, which includes zonal taxi crashes, network features, land-use patterns, temporal factors, and the essential data that constituted the exposure measures.

In 2012, the Planning Department of Hong Kong created a zoning system based on the Territory Survey of 2011. The resulting PDZ454 system divides the overall territory of Hong Kong into 398 zones. This zoning system was adopted in this study.

The Hong Kong Police Force and Transport Department collaboratively established a Traffic Information System to record detailed crash information (Wong et al., 2007). It includes the severity and environmental conditions (e.g., weather, lighting, and road type) for each crash. In 2011, there were 3,685 taxi-related crashes, of which 2,597 were multiple-vehicle crashes. These 2,597 multiple-vehicle crashes were used for model development in this case study. Six time periods were defined: 07:00 to 11:00 (morning), 11:00 to 15:00 (noon), 15:00 to 19:00 (afternoon), 19:00 to 23:00 (evening), 23:00 to 03:00 (midnight) and 03:00 to 07:00 (dawn) (Pei et al., 2012). Thus, a longitudinal cross-sectional panel data structure was applied in this case: the crashes were divided into 398 zones and 6 periods according to the location and time at which they occurred.

The road density, defined as the zonal road space R (i.e., zonal trafficable area) divided by the zonal area, and the intersection density, defined as the zonal intersection number divided by the zonal road space R, were anticipated as influential factors that contribute to crash risk. Vehicles interact with each other on road segments, and the interaction is even more intensive

416 at intersections, thus a higher zonal road density and a higher intersection density could
417 increase the likelihood of crashes. The zonal road space *R* and zonal area were extracted from
418 the digital map using ArcGIS.
419
420 A Traffic Characteristics Survey (TCS) conducted in Hong Kong in 2011 provided updated
421 travel data. The survey comprised three parts: a Household Interview Survey, a Stated
422 Preference Survey, and a Hotel/Guesthouse Tourists Survey (Transport Department, 2014).
423 Trip-destination information was extracted from the TCS database. Agglomerative
424 hierarchical cluster analysis was used to categorize zones according to land use, including
425 mainly residential areas, mainly workplace areas, residential and miscellaneous areas,
426 workplace and miscellaneous areas, retail areas, and cross-boundary areas (Meng et al., 2016).
427 The observed crash data and a summary of the contributory factors (including 2 continuous
428 variables and 10 dummy variables) are presented in Table 2.
429

Table 2. Summary of dependent and independent variables.

	Min.	Max.	Mean	Standard deviation (SD)
Dependent variables:				
No. of multiple-vehicle crashes	0	16	1.09	1.52
Continuous variables:				
Road density (%)	0.1	39.1	11.9	8.8
Intersection density (0.001*km^{-2})	0	3.04	0.26	0.26
Dummy variables:				
Land use				
Mainly residential area (baseline)			39.3%	
Mainly workplace area			10.4%	
Residential and miscellaneous area			22.1%	
Workplace and miscellaneous area			15.7%	
Retail area			12.4%	
Cross-boundary area			1.1%	
Time period				
03:00-07:00 (baseline)			16.6%	
07:00-11:00			16.7%	
11:00-15:00			16.7%	
15:00-19:00			16.7%	
19:00-23:00			16.7%	
23:00-03:00			16.6%	

The conventional taxi travel time exposure and the GDAE were the two candidate exposure measures chosen for this case study. The annual total traffic travel time T_{total} and the annual taxi travel time T_{taxi} of each zone in each time period were the essential ingredients of the two chosen exposure measures. However, these quantities were not observable. Therefore, linear data projection, which is a common data-scaling method that can estimate unobservable traffic data by multiplying the observable traffic data by a scaling factor (Wong and Wong, 2015, 2016a, 2016b), was used for the data estimation. The scaling factor is usually taken as a dimensionless ratio to bridge the observable traffic data and unobservable traffic data.

The *Annual Traffic Census* (ATC) 2011 (i.e., stationary source) and the taxi GPS database (i.e., mobile source) were used to constitute the observable traffic data and the corresponding scaling factors. The ATC report provided detailed traffic information, such as the annual average daily traffic data, obtained from 114 core stations and 730 coverage stations across Hong Kong (Lam et al., 2003; Tong, 2003). The core stations are distributed almost equally across the three districts of Hong Kong: 38 in Hong Kong Island, 33 in Kowloon, and 43 in the New Territories (Transport Department, 2012). Eighty-five core stations were chosen to represent the counting stations over the network. The AADT and the occupied probe taxi counts across each station were used to determine the scaling factors. The GPS data were obtained from GPS trackers installed in 460 probe taxis that traversed the Hong Kong network in 2011. The data comprised information on the taxi travel time, coordinates (in WGS84 format), speed, and direction at 30-second intervals. The travel times of the occupied probe taxis of each zone in each time period were the observable traffic data. These observations were obtained by multiplying the number of observed occupied taxi GPS records for each zone in each time period by 30 seconds. The total-traffic-to-probe-taxi ratio and total-taxi-to-probe-taxi ratio were the corresponding scaling factors used to estimate the annual total traffic travel time T_{total} and the annual taxi travel time T_{taxi}, respectively, using linear data projection. The scaling factors were estimated using the scaling factor estimation models proposed by Meng et al. (2016), which quantify the scaling factors as functions of a number of factors, such as the land use of a zone and distances between zones. Moreover, the zonal taxi average speed, $\bar{u}$, was estimated from the GPS data. Table 3 summarizes the logarithmic transformations of the conventional taxi travel time exposure, $\log(T_{taxi,it})$, and major factors of the GDAE, $\log(T_{taxi,it}T_{total,it})$, and $\log(I_{it})$, of the 398 zones and 6 time periods.

Table 3. Logarithmic transformations of taxi time exposure and factors of GDAE.

	Min.	Max.	Mean	SD
$\log(T_{taxi,it})$	9.40	9.57	9.48	0.044
$\log(T_{taxi,it}T_{total,it})$	20.7	21.0	20.8	0.081
$\log(I_{it})$	−5.86	7.34	−1.10	0.020

To prevent bias due to correlation and multicollinearity between the various independent variables, correlation tests and variance inflation factor tests of data associated with multiple-vehicle crashes were conducted. None of the independent variables in the dataset were found to be highly correlated with each other (all correlation figures were lower than 0.6), and all of the variance inflation factor values for the variables were less than 10. Therefore, there was no statistical evidence to suggest multicollinearity.

3.3 Results

To demonstrate the applicability of the proposed GDAE, zonal multiple-vehicle crash frequency models were calibrated on the basis of collected data. This subsection presents the results of the exposure measure selection, RPNB model establishment, and final modeling.

Two candidate exposure measures, conventional taxi travel time exposure and the proposed GDAE, were considered in this case study. If the conventional taxi travel time exposure is adopted, the crash frequency can be expressed as

$$\lambda_{it} = {T_{taxi,it}}^{\tau} \times \mathrm{EXP}(\boldsymbol{\beta}\mathbf{X}_{it} + \epsilon_i) \tag{11}$$

where $T_{taxi,it}$ is the annual taxi travel time for zone i in time period t; and τ is the model parameter that accounts for the nonlinear effect of the exposure measure. Previous studies have shown that the logarithmic transformation of an exposure measure could better fit the crash frequency function than the exposure measure itself (Kim and Washington, 2006; Mitra

and Washington, 2007; Washington et al, 2011). Therefore, the logarithmic form of annual zonal taxi travel time is used in this study.

In contrast, if the proposed GDAE is used, the crash frequency can be alternatively expressed as

$$\lambda_{it} = I_{it}^{\gamma_1}(T_{taxi,it}T_{total,it})^{\gamma_2} \times \mathrm{EXP}(\boldsymbol{\beta}\mathbf{X}_{it} + \epsilon_i), \quad (12)$$

where I_{it} is the state-topological factor in zone i in time period t and $T_{total,it}$ is the annual total traffic travel time in zone i in time period t.

To select the most representative exposure measure, two NB models were established for multiple-vehicle crashes involving taxis using only the candidate exposure measures. The conventional taxi travel time exposure was used in Model 1, and the GDAE measure was considered in Model 2. The results are presented in Table 4. A maximum likelihood estimation (MLE) approach was used to estimate the coefficients. The probability that each value of Z was above the upper limit or below the lower limit of the 95% confidence interval of the critical value is given as "P > $|\boldsymbol{z}|$" in the table.

As shown in Table 4, the AIC value of Model 2 is lower than that of Model 1; the MSE and RMSE values of the two models are quite similar, yet the predicted crash frequency of Model 2 (2599.4) is closer to the observed crash frequency (2597) than that of Model 1 (2589.8). The GDAE outperformed the conventional taxi travel time exposure, which provided statistical evidence that the GDAE should be a more representative exposure measure for modeling the multiple-vehicle crash frequency.

512 Table 4. NB models with only one exposure measure.

Variables	Model 1		Model 2	
	Coefficient	P > |z|	Coefficient	P > |z|
Constant	−3.467**	0.000	−4.242**	0.000
$\log(T_{taxi,it})$	0.352**	0.000	-	-
$\log(T_{taxi,it}T_{total,it})$	-	-	0.180**	0.000
$\log(I_{it})$	-	-	0.292**	0.000
Overdispersion parameter	0.545		0.485	
No. of observations	2385		2385	
Log likelihood	−3206.163		−3162.240	
AIC (df[a])	6418.326		6332.480	
MSE	1.888		1.889	
RMSE	1.374		1.374	
Predicted crash frequency	2589.8		2599.4	

513 a. df = degrees of freedom.

514 * Statistically significant at the 5% level.

515 ** Statistically significant at the 1% level.

516

517 The RPNB model was then calibrated for multiple-vehicle crashes involving taxis in Hong
518 Kong by further incorporating the collected explanatory variables that contributed to the
519 crash risk. A simulated MLE with 200 Halton draws was conducted (Train, 1999; Bhat,
520 2003). The normal distributions were used for all of the estimated coefficients, and the
521 coefficients with both a significant mean and standard deviation were considered to be
522 random, whereas the conventional fixed parameters were applied to the other coefficients.
523 Table 5 presents the results of the calibrated RPNB model.

524

525 It is worth noting that three variables in Table 5 had coefficients that were insignificant at the
526 5% level or above ("residential and miscellaneous area," "cross-boundary area," and
527 "intersection density"). To test the robustness and predictability of the model, these three
528 statistically insignificant variables were dropped, and the results turned out to be similar and
529 consistent with the model shown in Table 5 (i.e., all of the significant coefficients remained

530 significant and were very close to those presented in Table 5)[3]. To provide more
531 comprehensive information to the readers, these variables are presented in the final model.
532
533 Table 5. RPNB crash frequency models for multiple-vehicle crashes using GDAE.

Variables	Coefficient	Standard Error	P>|z|
Fixed parameters:			
Constant	−7.062**	0.311	0.000
$\log(T_{taxi,i}T_{total,i})$	0.134**	0.010	0.000
Residential and miscellaneous area	0.097	0.063	0.122
Retail area	0.465**	0.074	0.000
Cross-boundary area	−0.209	0.216	0.332
07:00-11:00	0.824**	0.089	0.000
11:00-15:00	0.671**	0.088	0.000
15:00-19:00	0.810**	0.086	0.000
19:00-23:00	0.777**	0.086	0.000
Intersection density in 0.001 (km^{-2})	−0.136	0.166	0.414
Road density (%)	2.539**	0.293	0.000
Means for random parameters:			
$\log(I_i)$	0.389**	0.037	0.000
Mainly workplace area	−0.456**	0.117	0.001
Workplace and miscellaneous area	0.298**	0.069	0.000
23:00-03:00	0.231**	0.093	0.013
Scale parameters for distributions of random parameters:			
$\log(I_i)$	0.017**	0.003	0.000
Mainly workplace area	0.678**	0.099	0.000
Workplace and miscellaneous area	0.253**	0.052	0.000
23:00-03:00	0.331**	0.060	0.000
Overdispersion Parameter	5.138**	0.869	0.000
No. of observations	2385		
Log likelihood at convergence	−3013.947		
Restricted log likelihood	−4476.931		
Pseudo R^2	0.327		
AIC	6067.894		
BIC	6183.423		

[3] The inclusion of the three insignificant variables may lead to a loss in efficiency, resulting in an increase in standard error of the estimated coefficients.

534 **3.4 Discussion**

535 This subsection discusses the results of the exposure measure selection and the calibrated
536 crash frequency model. A better understanding of the proposed GDAE and the identified
537 factors that contribute to multiple-vehicle crash risk could help improve taxi road safety.
538
539 According to the model calibration results of the RPNB model, 12 results (2 for the GDAE, 1
540 for the continuous explanatory variables, 8 for the subvariables of the categorical variables,
541 and 1 constant) were found to be significant at the 5% level or above, among which eight
542 were fixed parameters and four were random variables. These results reveal that unobserved
543 heterogeneity across various zones existed in the crash frequency model regarding taxis in
544 Hong Kong and were captured by the four random parameters acquired in the RPNB model.
545
546 For the proposed GDAE, both correction terms were significant at the 5% level, in which γ_1
547 was fixed (coefficient, 0.134) and γ_2 was random (with a mean of 0.389 and scale parameter
548 0.017). Thus, heterogeneity existed in the state-topological factor, I_i. Because the same zonal
549 average speed under different traffic volumes may result in different vehicular meeting
550 potentials, the effect of the state-topological factor was intuitively heterogeneous and thus
551 resulted in the heterogeneous effect of GDAE on the multiple-vehicle crash frequency. Based
552 on the calibrated distribution of γ_2, the 95% confidence interval was between 0.292 and
553 0.496. Because the lower boundary was greater than 0 and the upper boundary was less than
554 1, we have sufficient confidence to believe that both the travel time effect, $T_{taxi,i}T_{total,i}$, and
555 the state-topological factor, I_i, had positive effects on the multiple-vehicle crash rate.
556 Moreover, the growth rate of GDAE decreased with $T_{taxi,i}T_{total,i}$ and I_i, indicating that the
557 increase in the meeting frequency of taxis and other vehicles due to the increase in
558 $T_{taxi,i}T_{total,i}$ and I_i became less effective as their values increased.

559

560 Three explanatory variables had random coefficients on the zonal risk of multiple-vehicle
561 crashes involving taxis, namely "mainly workplace area," "workplace and miscellaneous
562 area," and "23:00 to 3:00". The heterogeneity effects of these covariates are discussed in
563 relation to the other land use types and time periods below.

564

565 To study the effects of the time periods on the crash risk, 03:00 to 07:00 (i.e., dawn) was
566 selected as the baseline reference. Compared to the baseline time period, 07:00 to 11:00
567 (coefficient, 0.824), including the morning peak hours, and 15:00 to 19:00 (coefficient, 0.810)
568 and 19:00 to 23:00 (coefficient, 0.777), covering the afternoon and evening peak hours, were
569 the three most dangerous periods of the day. (It should be noted that traffic in Hong Kong
570 during the evening is usually still considered "busy".) Obviously, the number of passengers
571 and the intensity of taxi activity were the highest in the morning and afternoon peak hours,
572 especially on weekdays. During these busy hours, taxi drivers must concentrate on activities
573 such as cruising, searching for passengers, and picking up and dropping off passengers. This
574 heavy workload could possibly lead to driver fatigue, which could make the taxi drivers less
575 aware of possible dangerous situations. Thus, they might not be able to respond sufficiently
576 quickly to avoid crashes. The period from 11:00 to 15:00 (coefficient, 0.671), here referred to
577 as "noon," was found to be less risky than the morning and afternoon peaks, because the
578 workload of the taxi drivers during this period was relatively low. Compared with 03:00 to
579 07:00, the effect of 23:00 to 03:00 was heterogeneous, and the 95% confidence interval of the
580 random coefficient for 23:00 to 03:00 was between −0.344 to 0.806. This heterogeneous
581 effect could have resulted from the highly uneven spatial distribution of the taxis due to their
582 special cruising behavior, which was influenced by the time-specific attractions in the zones
583 with intensive night activities.

584

585 The land use categorical explanatory variable “mainly residential area” was chosen as the
586 baseline reference. The highly significant calibrated mean of the random coefficient for
587 “mainly workplace areas” (−0.456) showed that the risk of multiple-vehicle crashes involving
588 taxis in those areas was lower than that of “mainly residential areas” for most cases, yet the
589 scale parameter (0.678) indicated that there were exceptions. Although workplaces normally
590 attract intense traffic, the intensity of the attraction varies with the location of the workplace.
591 In Hong Kong, workplaces are concentrated in commercial and administrative areas, such as
592 Central and Admiralty. The traffic density is extremely high in these areas, especially during
593 workdays. Compared with residential areas, workplace areas have heavier traffic and attract
594 more taxi trips, which make taxi drivers considerably more cautious when driving in these
595 areas. In the New Territories, however, zones with large industrial areas are also categorized
596 as workplaces, but the traffic is relatively lighter with a certain number of taxi trips.
597 Compared with some residential areas, the multiple-vehicle crash risk in such areas is likely
598 to be lower. Moreover, compared to “mainly residential areas,” “retail areas” were associated
599 with a higher multiple-vehicle crash risk (coefficient, 0.465). The total zonal crash risk has
600 been shown to be higher in mixed land-use zones than in any other land-use type (Pulugurtha
601 et al., 2013; Chen, 2015). Because the land use proportions of the retail areas sampled in this
602 study were similar to those of the mixed land-use zones, our finding of the effects of “retail
603 areas” on the crash risk is generally consistent with those of previous studies.

604

605 The zonal road density, obtained by dividing the zonal road space by the zonal area, had a
606 positive effect on the crash risk (coefficient, 2.539). Vehicles interact on the road space, and
607 in certain circumstances, some interactions result in crashes. Given the same zonal area, a

608 zone with greater zonal density offers more road space for interactions among vehicles and
609 hence leads to a greater crash risk.
610
611 **4 Conclusions**
612 This study proposes a more representative exposure measure for modeling multiple-vehicle
613 crash frequency. We analogized the meeting frequency of vehicles with the meeting
614 frequency of gas molecules. Based on the central idea of the classical collision theory in
615 physical chemistry, the meeting frequency function of vehicles was derived. It was found that
616 the meeting frequency of vehicles is, theoretically, dependent on the time exposures of the
617 two vehicle types of interest, the mean speed of the vehicles, the road space of a given area,
618 the effective meeting width, and the angle of the vehicles. However, at the current stage,
619 direct application of the meeting frequency function may not be possible because the
620 effective meeting width and the angle of the vehicles—two unknown constants—of a study
621 region of interest are not easily obtainable. Thus, the GDAE was formulated by means of the
622 obtainable major factors identified in the meeting frequency function. Correction terms were
623 incorporated to account for any differences between the idealized scenario and reality.
624 Compared to conventional exposure measures, the proposed GDAE can provide further
625 insight into the physics of the vehicle meeting mechanism, which is its major distinctive
626 feature and the major contribution of this study.
627
628 To provide statistical evidence on the applicability of the proposed GDAE, a zonal multiple-
629 vehicle crash frequency model involving taxis and total traffic as the two chosen vehicle
630 types was established on the basis of the crash data from Hong Kong in 2011. The
631 performance of the GDAE was compared with that of the conventional time exposure in
632 modeling multiple-vehicle crashes involving taxis. The GDAE was found to be a better

633 exposure measure of multiple-vehicle crash frequency than the conventional time exposure
634 based on the information criterion.
635
636 The explanatory factors that contributed significantly to the crash risk of taxis and total traffic
637 were then identified on the basis of an RPNB model that addressed the possible unobserved
638 heterogeneity. The state-topological factor was found to have a heterogeneous effect on the
639 multiple-vehicle crash risk involving taxis, whereas the travel time measurement had a fixed
640 positive effect. The relatively busy periods in Hong Kong (i.e., 07:00 to 11:00, 15:00 to 19:00,
641 and 19:00 to 23:00) were found to be the most dangerous times of day in terms of the
642 likelihood of multiple-vehicle crashes involving taxis. In terms of land use, "retail area" was
643 the riskiest of the different land-use areas. Furthermore, the crash risk was found to increase
644 with the road density.
645
646 The proposed GDAE is a novel and promising proxy exposure measure for modeling
647 multiple-vehicle crash frequency. With a more representative exposure measure, it can
648 facilitate the identification of factors that contribute to crash risk and hence the formulation of
649 policies to improve road safety. Further incorporation of the effective meeting width and
650 angle in the exposure measure and the application of this proxy to less aggregated datasets
651 present interesting future research directions. Moreover, the GDAE is derived from the
652 meeting frequency function, which quantifies the number of potential traffic conflicts.
653 Because traffic conflict is an essential and important concept in transportation research, the
654 proposed GDAE could be used in other cases with suitable modifications. In addition, due to
655 the limitations of the dataset available, the empirical modeling of taxi crashes in Hong Kong
656 lacks contributory factors such as roadway geometric characteristics and environmental
657 conditions. Future studies may investigate the effects of these variables with crash data

neutralized by GDAE and explore other modeling approaches, such as multivariate, latent-class, zero-inflated, and space-time models with incorporation of GDAE.

Acknowledgments

This study was supported by a Research Postgraduate Studentship, grants from the University Research Committee of the University of Hong Kong (201511159015), and the Joint Research Scheme of the National Natural Science Foundation of China/Research Grants Council of Hong Kong (Project Nos. 71561167001 and N_HKU707/15). The third author was also supported by the Francis S. Y. Bong Professorship in Engineering. We express our special thanks to Concord Pacific Satellite Technologies Ltd. and Motion Power Media Ltd. for providing the GPS taxi data and to the Transport Department of the Hong Kong Special Administrative Region for providing the TCS and ATC data.

References

Amoh-Gyimah, R., Saberi, M., Sarvi, M., 2017. The effect of variations in spatial units on unobserved heterogeneity in macroscopic crash models. Analytic Methods in Accident Research 13, 28-51.

Anastasopoulos, P., Mannering, F., 2009. A note on modeling vehicle accident frequencies with random-parameters count models. Accident Analysis and Prevention 41 (1), 153-159.

Baker, S.P., Wong, J., Baron, R.D., 1976. Professional drivers: Protection needed for a high-risk occupation. American Journal of Public Health 66 (7), 649-654.

682 Barua, S., El-Basyouny, K., Islam, M.T., 2015. Effects of spatial correlation in random
683 parameters collision count-data models. Analytic Methods in Accident Research 5-6,
684 28-42.
685 Barua, S., El-Basyouny, K., Islam, M.T., 2016. Multivariate random parameters collision
686 count data models with spatial heterogeneity. Analytic Methods in Accident Research
687 9, 1-15.
688 Bhat, C., 2003. Simulation estimation of mixed discrete choice models using randomized and
689 scrambled Halton sequences. Transportation Research Part B: Methodological 37 (1),
690 837-855.
691 Bhat, C., Born, K., Sidharthan, R., Bhat, P., 2014. A count data model with endogenous
692 covariates: formulation and application to roadway crash frequency at intersections.
693 Analytic Methods in Accident Research 1, 53-71.
694 Bie, J., Lo, H.K., Wong, S., Hung, W., Loo, B.P., 2005. Safety analysis of traffic roundabout:
695 Conventional versus Alberta-type markings. Journal of the Eastern Asia Society for
696 Transportation Studies 6, 3309-3324.
697 Buddhavarapu, P., Scott J. G., Prozzi, J.A., 2016. Modeling unobserved heterogeneity using
698 finite mixture random parameters for spatially correlated discrete count data.
699 Transportation Research Part B 91, 492-510.
700 Burns, P.C., Wilde, G.J.S., 1995. Risk taking in male taxi drivers: Relationships among
701 personality, observational data and driver records. Personality and Individual
702 Differences 18 (2), 267-278.
703 Chapman, R., 1973. The concept of exposure. Accident Analysis and Prevention 5 (2), 95-
704 110.

705 Chen, E., Tarko, A.P., 2014. Modeling safety of highway work zones with random
706 parameters and random effects models. Analytic Methods in Accident Research 1, 86-
707 95.
708 Chen, P., 2015. Built environment factors in explaining the automobile-involved bicycle
709 crash frequencies: A spatial statistic approach. Safety Science 79, 336-343.
710 Chen, C., Xie, Y., 2016. Modeling the effects of AADT on predicting multiple-vehicle
711 crashes at urban and suburban signalized intersections. Accident Analysis and
712 Prevention 91, 72-83.
713 Chiou, Y-C., Fu, C., Chih-Wei, H., 2014. Incorporating spatial dependence in simultaneously
714 modeling crash frequency and severity. Analytic Methods in Accident Research 2, 1-
715 11.
716 Chiou, Y-C., Fu, C., 2015. Modeling crash frequency and severity with spatiotemporal
717 dependence. Analytic Methods in Accident Research 5-6, 43-58.
718 Chipman, M.L., MacGregor, C.G., Smiley, A.M., Lee-Gosselin, M., 1993. The role of
719 exposure in comparisons of crash risk among different drivers and driving
720 environments. Accident Analysis and Prevention 25 (2), 207-211.
721 Coruh, E., Bilgic, A., Tortum, A., 2015. Accident analysis with aggregated data: The random
722 parameters negative binomial panel count data model. Analytic Methods in Accident
723 Research 7, 37-49.
724 Dalziel, J.R., Job, R.F.S., 1997. Motor vehicle accidents, fatigue and optimism bias in taxi
725 drivers. Accident Analysis and Prevention 29 (4), 489-494.
726 Elvik, R., 2015. Some implications of an event-based definition of exposure to the risk of
727 road accident. Accident Analysis and Prevention 76, 15-24.

728 Fridstrøm, L., Ifver, J., Ingebrigtsen, S., Kulmala, R., Thomsen, L.K., 1995. Measuring the
729 contribution of randomness, exposure, weather, and daylight to the variation in road
730 accident counts. Accident Analysis and Prevention 27 (1), 1-20.
731 Greene, W., 2007. Limdep, Version 9.0. Econometric Software Inc., Plainview, NY.
732 Heydari, S., Fu, L., Miranda-Moreno, L.F., Joseph, L., 2017. Using a flexible multivariate
733 latent class approach to model correlated outcomes: A joint analysis of pedestrian and
734 cyclist injuries. Analytic Methods in Accident Research 13, 16-27.
735 Huang, H., Chin, H.C., 2009. Disaggregate propensity study on red light running crashes
736 using quasi-induced exposure method. Journal of Transportation Engineering 135 (3),
737 104-111.
738 Huang, H., Zhou, H., Wang, J., Chang, F., Ma, M., 2017. A multivariate spatial model of
739 crash frequency by transportation modes for urban intersections. Analytic Methods in
740 Accident Research 14, 10-21.
741 Imprialou, M.M., Quddus, M., Pitfield, D.E., 2016. Predicting the safety impact of a speed
742 limit increase using condition-based multivariate Poisson lognormal regression.
743 Transportation Planning and Technology 39 (1), 3-23.
744 Jiang, X., Lyles, R.W., Guo, R., 2014. A comprehensive review on the quasi-induced
745 exposure technique. Accident Analysis and Prevention 65, 36-46.
746 Johnson, N.J., Sorlie, P.D., Backlund, E., 1999. The impact of specific occupation on
747 mortality in the US national longitudinal mortality study. Demography 36 (3), 355-
748 367.
749 Kim, D.G., Washington, S., 2006. The significance of endogeneity problems in crash models:
750 An examination of left-turn lanes in intersection crash models. Accident Analysis and
751 Prevention 38 (6), 1094-100.

752 Kroyer, H.R.G., 2016. Pedestrian and bicyclist flows in accident modeling at intersections:
753 Influence of the length of observational period. Safety Science 82. 315-324.
754 La, Q.N., Lee, A.H., Meuleners, L.B., Van Duong, D., 2013. Prevalence and factors
755 associated with road traffic crash among taxi drivers in Hanoi, Vietnam Accident
756 Analysis and Prevention 50, 451-455.
757 Laidler, K.J., 1973. Chemical Kinetics, 3rd edition. Tata McGraw-Hill, US.
758 Lam, L.T., 2004. Environmental factors associated with crash-related mortality and injury
759 among taxi drivers in New South Wales, Australia. Accident Analysis and Prevention
760 36 (5), 905-908.
761 Lam, W.H.K., Hung, W.T., Lo, H.K., Lo, H.P., Tong, C.O., Wong, S.C., Yang, H., 2003.
762 Advancement of the annual traffic census in Hong Kong. In: Proceedings of the
763 Institution of Civil Engineers-Transport 156 (2), 103-115.
764 Lee, J., Abdel-Aty, M., Jiang, X., 2015. Multivariate crash modeling for motor vehicle and
765 non-motorized modes at the macroscopic level. Accident Analysis and Prevention 78,
766 146-154.
767 Li, G., Braver, E.R., Chen, L., 2003. Fragility versus excessive crash involvement as
768 determinants of high death rates per vehicle-mile of travel among old drivers.
769 Accident Analysis and Prevention 35, 227–235.
770 Machin, M.A., De Souza, J.M.D., 2004. Predicting health outcomes and safety behaviour in
771 taxi drivers. Transportation Research Part F: Traffic Psychology and Behaviour 7 (4-
772 5), 257-270.
773 Mannering, F.L., Shankar, V., Bhat, C.R., 2016. Unobserved heterogeneity and the statistical
774 analysis of highway accident data. Analytic Methods in Accident Research 11, 1-16.

775 Meng, F., Wong, S.C., Wong, W., Li, Y. C., 2017. Estimation of scaling factors for traffic
776 counts based on stationary and mobile sources of data. International Journal of
777 Intelligent Transportation Systems Research 15 (3), 180-191.
778 Miranda-Moreno, L., Strauss, J., Morency, P., 2011. Disaggregate exposure measures and
779 injury frequency models of cyclist safety at signalized intersections. Transportation
780 Research Record: Journal of the Transportation Research Board 2236, 74-82.
781 Mitra, S., Washington, S., 2007. On the nature of over-dispersion in motor vehicle crash
782 prediction models. Accident Analysis and Prevention 39 (3), 459-468.
783 Naci, H., Chisholm, D., Baker, T.D., 2009. Distribution of road traffic deaths by road user
784 group: a global comparison. Injury Prevention 15, 55-59.
785 Pei, X., Wong, S.C., Sze, N.N., 2012. The roles of exposure and speed in road safety analysis.
786 Accident Analysis and Prevention 48, 464-471.
787 Pei, X., Sze, N.N., Wong, S.C., Yao, D., 2016. Bootstrap resampling approach to
788 disaggregated analysis of road crashes in Hong Kong. Accident Analysis and
789 Prevention 95, 512-520.
790 Pulugurtha, S.S., Duddu, V.R., Kotagiri, Y., 2013. Traffic analysis zone level crash
791 estimation models based on land use characteristics. Accident Analysis and
792 Prevention 50, 678-687.
793 Qin, X., Ivan, J.N., Ravishanker, N., 2004. Selecting exposure measures in crash rate
794 prediction for two-lane highway segments. Accident Analysis and Prevention 36, 183-
795 191.
796 Qin, X., Ivan, J.N., Ravishanker, N., Liu, J., Tepas, D., 2006. Bayesian estimation of hourly
797 exposure functions by crash type and time of day. Accident Analysis and Prevention
798 38 (6), 1071-1080.

799 Rosenbloom, T., Shahar, A., 2007. Differences between taxi and nonprofessional male
800 drivers in attitudes towards traffic-violation penalties. Transportation Research Part F:
801 Traffic Psychology and Behaviour 10 (5), 428-435.
802 Russo, B.J., Savolainen, P.T., Schneider, W.H. IV, Anastasopoulos, P., 2014. Comparison of
803 factors affecting injury severity in angle collisions by fault status using a random
804 parameters bivariate ordered probit model. Analytic Methods in Accident Research 2,
805 21-29.
806 Sarwar, M.T., Anastasopoulos, P., 2017. The effect of long term non-invasive pavement
807 deterioration on accident injury-severity rates: a seemingly unrelated and multivariate
808 equations approach. Analytic Methods in Accident Research 13, 1-15.
809 Serhiyenko, V., Mamun, S.A., Ivan, J.N., Ravishanker, N., 2016. Fast Bayesian inference for
810 modeling multivariate crash counts. Analytic Methods in Accident Research 9, 44-53.
811 Shams, M., Shojaeizadeh, D., Majdzadeh, R., Rashidian, A., Montazeri, A., 2011. Taxi
812 drivers' views on risky driving behavior in Tehran: A qualitative study using a social
813 marketing approach. Accident Analysis and Prevention 43 (3), 646-51.
814 Stamatiadis, N., Deacon, J.A., 1997. Quasi-induced exposure: Methodology and insight.
815 Accident Analysis and Prevention 29 (1), 37-52.
816 Sullman, M.J., Stephens, A.N., Kuzu, D., 2013. The expression of anger amongst Turkish
817 taxi drivers. Accident Analysis and Prevention 56, 42-50.
818 Tong, C.O., Hung, W. T., Lam, W. H. K., Lo, H.P., Wong, S.C., Yang, H., 2003. A new
819 survey methodology for the annual traffic census in Hong Kong. Traffic Engineering
820 and Control 44, 214-218.
821 Transport Department, 2012. The Annual Traffic Census 2011. Transport Department,
822 HKSAR.

Transport Department, 2014. Road Traffic Accident Statistics. Transport Department, HKSAR.

Transport Department, 2014. Travel Characteristics Survey 2011 Final Report. Transport Department, HKSAR.

Train, K., 1999. Halton Sequences for Mixed Logit. Working Paper. University of California, Department of Economics, Berkley.

Tulu, G.S., Washington, S., Haque, M.M., King, M.J., 2015. Investigation of pedestrian crashes on two-way two-lane rural roads in Ethiopia. Accident Analysis and Prevention 78, 118-126.

Venkataraman, N., Ulfarsson, G., Shankar, V., Oh, J., Park, M., 2011. Model of relationship between interstate crash occurrence and geometrics: Explanatory insights from random parameter negative binomial approach. Transportation Research Record 2236, 41-48.

Venkataraman, N., Ulfarsson, G.F., Shankar, V.N., 2013. Random parameter models of interstate crash frequencies by severity number of vehicles involved, collision and location type. Accident Analysis and Prevention 50, 309-318.

Washington, S.P., Karlaftis, M.G., Mannering, F.L., 2011. Statistical and econometric methods for transportation data analysis (Second Edition). New York: CRC Press.

Wolfe, A.C., 1982. The concept of exposure to the risk of a road traffic accident and an overview of exposure data collection methods. Accident Analysis and Prevention 14 (5), 337-340.

Wong, S.C., Sze, N.N., Li, Y.C., 2007. Contributory factors to traffic crashes at signalized intersections in Hong Kong. Accident Analysis and Prevention 39 (6), 1107-1113.

Wong, W., Wong, S.C., 2015. Systematic bias in transport model calibration arising from the variability of linear data projection. Transportation Research Part B: Methodological 75, 1-18.

Wong, W., Wong, S.C., 2016a. Biased standard error estimations in transport model calibration due to heteroscedasticity arising from the variability of linear data projection. Transportation Research Part B: Methodological 88, 72-92.

Wong, W., Wong, S.C., 2016b. Unbiased calibration of nonlinear transport models based on linearly projected data: A case study of macroscopic fundamental diagram. Transportation Science. Under review.

Xiong, Y., Mannering, F., 2013. The heteroscedastic effects of guardian supervision on adolescent driver-injury severities: A finite mixture-random parameters approach. Transportation Research Part B 49, 39-54.